AD-A096 724 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 5/8
BIASED RESULTS WITH THE LEAVING-ONE-OUT METHOD OF PATTERN RECOG--ETC(U)
DEC 80 L K JONES, K H WICKWIRE F19628-80-C-0002
UNCLASSIFIED TR-545 ESD-TR-80-233 NL

END
DATE FILMED
4-81
DTIC

ESD-TR-80-233

LEVEL II

12

SC



# Technical Report

545

L.K. Jones

K.H. Wickwire

## Biased Results with the Leaving-One-Out Method of Pattern Recognition



11 December 1980

Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-80-C-0002 by

**Lincoln Laboratory**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

*Lexington, Massachusetts*

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-80-C-0002.





This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle, Lt.Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

# BIASED RESULTS WITH THE LEAVING-ONE-OUT METHOD OF PATTERN RECOGNITION.

*L.K. JONES*
*K.H. WICKWIRE*
*Group 92*



TECHNICAL REPORT 545

11 DECEMBER 1980



LEXINGTON MASSACHUSETTS

## ABSTRACT

We show that the leaving-one-out method of pattern recognition must yield biased results when the two sets of training data (representing two classes to be discriminated) are identical. This phenomenon, which we observed during a study of the sensitivity of classification results to errors in the training data, can be eliminated by generating the training sets independently.

## 1. Introduction

In certain studies of the sensitivity of classification results to errors in the data submitted for classification, one submits for training two sets of data with the property that the second set has been derived from the first by corrupting (adding noise or bias to) the first set. An example of such a study arises in the assessment of the effect of measurement noise on the accuracy of radar static patterns from reentry vehicles. A procedure for assessing this effect is to add noise of increasing magnitude to a static pattern defined to be free of noise and then submit signatures derived from the noisy and noise-free patterns to a classification algorithm. So long as the "noisy" signatures cannot be discriminated (according to some probability of error criterion) from the noise-free signatures, the noise in the static patterns is accepted. Noise levels which lead to signatures which can be discriminated from those corresponding to a noise-free pattern are not accepted. The leaving-one-out (L-) method of classification has properties which make it suitable for these and other investigations: it makes efficient use of the data, which is desirable when there are not many available, and its expected error is an upper bound on the Bayes classification error. (For details, see [1], Chap. 6.). It is the purpose of this note to justify the following recommendation: if the L-method is used to perform discrimination in

such studies, then the corrupted data should be generated not from the uncorrupted data, but from a set of data which is statistically (but not identically) equivalent to it. The reason for this recommendation is the fact (to be demonstrated below) that the L-method produces strongly biased results when it is exercised upon data sets which are identical, or upon those with the property that one set is a (slight) corruption of the other.

## 2. The biasness of the L-method when classes are represented by identical data sets

The plan of this section is to show first, that when the two classes described above have a one-dimensional distribution, the Gaussian density estimate for the class from which a sample has been left out is smaller than the Gaussian density estimate for the class which includes the sample. (We shall exclude throughout the trivial case in which all observations in the full sample are identical.) Since reasonable classifiers are based upon the comparison (for example, through (log-) likelihood ratios) of density estimators evaluated at points in the test set, the biasness of the L-method follows in this case. Next we show, using the previous result, that the L-method is biased in these circumstances for classes from a multi-dimensional distribution when the multi-variate Gaussian density estimator is used. Finally, we motivate the result for the case of a general density estimate by proving

it true for the Parzen estimator. We expect that a similar argument can be found to prove L-method biasness for identical data sets in the case of other reasonable density estimators, but we shall not do that here.

A. Classes drawn from a one-dimensional distribution

Let $x_1, \ldots, x_{N-1}, x_N$ be the variables in the class from which $x_N$ is left out during the training (estimation) part of classifier design. Since the performance of the L-method is invariant with respect to shifts of the data, we may assume that $(N-1)^{-1} \sum_{i=1}^{N-1} x_i = 0$.

There are then three cases to consider:

Case A1. $x_1 = x_2 = \cdots = x_{N-1}$, $x_N \neq 0$. The contribution to the Gaussian estimate at $x_N$ is greater than zero, so that the estimate based on $x_1, \ldots, x_{N-1}, x_N$ is greater than that based on $x_1, \ldots, x_{N-1}$.

Let $s_N^2 \equiv (1/N) \sum_{i=1}^{N} x_i^2$ and assume for the next two cases, without loss of generality, that $s_{N-1}^2 = 1$.

Case A2. $x_N = 0$. The conclusion follows from the fact that $s_N < s_{N-1}$, which implies that the N-point density estimate is larger than the (N-1)-point density estimate.

Case A3. $x_N \neq 0$. The comparable terms in the (N-1)-and N-point density estimates are $\exp(-x_N^2/2)$ and

$$\left[\frac{N-1}{N} + x_N^2\left(\frac{N-1}{N^2}\right)\right]^{-\frac{1}{2}} \cdot \exp\left\{-\frac{x_N^2}{2}\left(\frac{N-1}{N}\right)^2 \Big/ \left(\frac{N-1}{N} + x_N^2\left(\frac{N-1}{N^2}\right)\right)\right\} ,$$

and we must show that the latter is greater than the former. There are two cases to consider.

Case A3i. $(N-1)/N + x_N^2(N-1)/N^2 \leq 1$ . Here it is enough to show that the exponent of the latter is smaller than that of the former. This follows from

$$\frac{(x_N - x_N/N)^2}{\frac{N-1}{N} + \frac{x_N^2}{N} - \frac{x_N^2}{N^2}} = \frac{x_N^2\left(\frac{N-1}{N}\right)^2}{\frac{N-1}{N} + \frac{x_N^2}{N}\left(\frac{N-1}{N}\right)}$$

$$= \frac{x_N^2\left(\frac{N-1}{N}\right)^2}{\frac{N-1}{N} + \frac{x_N^2}{N-1}\left(\frac{N-1}{N}\right)^2} = \frac{x_N^2}{\frac{N}{N-1} + \frac{x_N^2}{N-1}} < x_N^2 ,$$

whence the biasness follows.

Case A3ii. $(N-1)/N + x_N^2(N-1)/N^2 > 1$ . Here it is enough to show that

$$\left[\frac{N-1}{N} + x_N^2\left(\frac{N-1}{N^2}\right)\right]^{-\frac{1}{2}} \exp\left\{-\frac{x_N^2}{2}\left(\frac{N-1}{N}\right)^2\right\} \geq \exp\left(-x_N^2/2\right) .$$

This is the same as showing that

$$\frac{x_N^2}{2} - \tfrac{1}{2}\log\left[\frac{N-1}{N} + x_N^2\left(\frac{N-1}{N^2}\right)\right] - \frac{x_N^2}{2}\left(\frac{N-1}{N}\right)^2 \geq 0.$$

By the monotonicity of the logarithm, the left side of this inequality is greater than

$$\frac{x_N^2}{2}\left(\frac{2N-1}{N^2}\right) - \tfrac{1}{2}\log\left[1 + x_N^2\left(\frac{N-1}{N^2}\right)\right] ,$$

which by the same monotonicity is greater than

$$\frac{x_N^2}{2}\left(\frac{2N-1}{N^2}\right) - \tfrac{1}{2}\log\left[1 + x_N^2\left(\frac{2N-1}{N^2}\right)\right] .$$

But this is of the form $\frac{1}{2}\left(\mu - \log(1+\mu)\right)$, which is positive for $\mu > 0$, and the biasness follows in this case as well.

B. Classes drawn from a multi-dimensional distribution

Let the estimates of the covariance matrices of the two design classes be denoted by $\hat{\Sigma}_i$, i=1,2. One can always find a linear transformation of the data which simultaneously diagonalizes $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ (even if one or both are singular). The Gaussian density estimator for the transformed data is the product of one-dimensional Gaussian densities and we can therefore appeal to the result in the one-dimensional case. Suppose that X is the M-vector deleted by the L-method and A is the set of vectors remaining in the corresponding class. Let

$\hat{P}(B)$ be the Gaussian estimate corresponding to a data set B and $\hat{P}^i(B)$ the projection of $\hat{P}(B)$ onto its i-th component. Let $Y = (y_1, \ldots, y_M)$ be any vector in A and put $A^i \equiv \{y_i : Y \in A\}$. Consider any component $x_i$ of X and note that if $A^i = \{x_i\}$, then $\hat{P}^i(A) = \hat{P}^i(A \cup \{X\})$, and if $A^i \neq \{x_i\}$, then $\hat{P}^i(A) < \hat{P}^i(A \cup \{X\})$ at $x_i$. Since the data are not identical by assumption, there exists at least one i such that $A^i \neq \{x_i\}$. Hence, at least one of the factors of $\hat{P}(A)$ is smaller than the corresponding factor of $\hat{P}(A \cup \{X\})$, whence $\hat{P}(A) < \hat{P}(A \cup \{X\})$ at X.

C. Classes drawn from an arbitrary distribution; the result motivated by consideration of the Parzen density estimator.

In the general case, the underlying density cannot be summarized by its first-and second-order moments, and it is not easy to choose a density estimator which is suitable for all reasonable classification procedures. One estimator which has earned wide acceptance because of its attractive properties (asymptotic unbiasedness and consistency) is the Parzen estimator. For the estimation of one-dimensional densities by the Parzen technique one usually chooses a kernel function which has a unique maximum at its central point. The Gaussian kernel has this property in any number of dimensions. We shall prove below L-method biasness for data whose densities are estimated using the Gaussian kernel or any other multi-dimensional kernel which has a unique maximum at its central point.

Suppose that the class in question is represented by N-1 data points $x_1, \ldots, x_{N-1}$ plus the point $x_N$, which is deleted during estimation by the L-method. Let $\hat{P}_N(x) = (1/N)\sum_{i=1}^{N} p(x_i;x)$ be the N-point density estimate evaluated at x, where $p(x_i;\cdot)$ is the estimation kernel (cf. [1], Chap. 6). Then

$$\hat{P}_N(\cdot) = \frac{N-1}{N}\,\hat{P}_{N-1}(\cdot) + \frac{1}{N}\,p(x_N;\cdot)$$

and at $x_N$,

$$\hat{P}_N(x_N) > \frac{N-1}{N}\,\hat{P}_{N-1}(x_N) + \frac{1}{N}\,\hat{P}_{N-1}(x_N) = \hat{P}_{N-1}(x_N)$$

since $\hat{P}_{N-1}(x_N) < p(x_N;x_N)$. The last inequality follows from the fact that $\hat{P}_{N-1}(x_N)$ is the mean of N-1 terms $p(x_i;x_N) \leq p(x_N;x_N)$, and at least one of these terms is strictly less than $p(x_N;x_N)$. Hence, the Parzen estimate for a set of N data points is greater than that obtained from any (N-1)-point subset of them, which was to be proved.

## 3. Conclusions and remarks

We have shown that if the leaving-one-out method is employed to discriminate a set of data points from one which is identical to it, then biased results will always be obtained if the Gaussian density estimator is used. In particular, the leakage and false alarm rates will be 100% in this case.

If other density estimators are used, we expect the same conclusion to hold and we motivate our expectation by consideration of distributions whose densities are estimated using the Parzen estimator. If noise of "small" magnitude is added subsequently to the data, or if noise of "moderate" magnitude is added to a small number of the components of the data, one can see that the inequality of density estimates proved in Section 2 is often preserved, so that the L-method will still yield biased results. To remove this L-method bias one should generate the two classes independently, so that one starts with classes which are statistically, but not identically equivalent. Experiments with a quadratic classifier in which this procedure was applied to simulated Gaussian data and to non-Gaussian data from a sensitivity study confirm that independent generation does remove the bias of the L-method.

## REFERENCES

[1]. K. Fukunaga, Introduction to Statistical Pattern Recognition (Academic Press, New York, 1972).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>ESD-TR-80-233 | 2. GOVT ACCESSION NO.<br>AD-A096 124 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE *(and Subtitle)*<br>Biased Results with the Leaving-One-Out Method of Pattern Recognition | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical Report<br>6. PERFORMING ORG. REPORT NUMBER<br>Technical Report 545 |
| 7. AUTHOR*(s)*<br>Lee K. Jones and Kenneth H. Wickwire | | 8. CONTRACT OR GRANT NUMBER*(s)*<br>F19628-80-C-0002 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Lincoln Laboratory, M.I.T.<br>P.O. Box 73<br>Lexington, MA 02173 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>Program Element No. 63311F<br>Project No. 627A |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Air Force Systems Command, USAF<br>Andrews AFB<br>Washington, DC 20331 | | 12. REPORT DATE<br>11 December 1980<br>13. NUMBER OF PAGES<br>14 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*<br>Electronic Systems Division<br>Hanscom AFB<br>Bedford, MA 01731 | | 15. SECURITY CLASS. *(of this report)*<br>Unclassified<br>15a. DECLASSIFICATION DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

statical pattern recognition

biasness of the leaving-one-out method

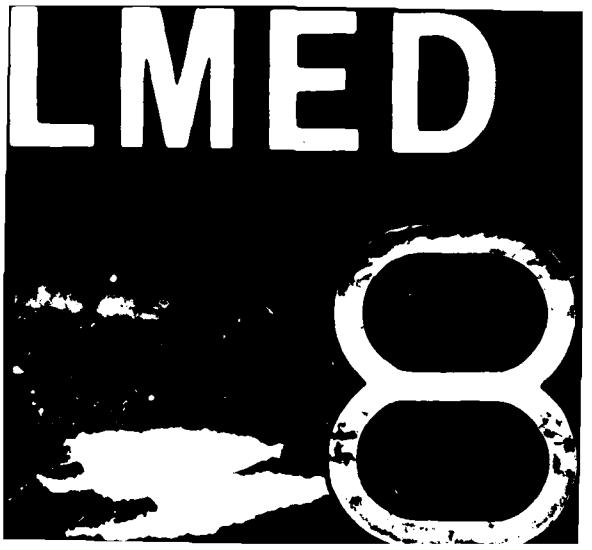
20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

We show that the leaving-one-out method of pattern recognition must yield biased results when the two sets of training data (representing two classes to be discriminated) are identical. This phenomenon, which we observed during a study of the sensitivity of classification results to errors in the training data, can be eliminated by generating the training sets independently.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*


LMED
8